

CSE842: Natural Language Processing

Lecture 15: Computational Lexical Semantics

3/4/2009

CSE842, Spring 2009, MSU

1

- The Final Project Guideline has been posted on Angel.
- You are encouraged to come to see me (if you have any questions) before March 30 (when the proposal is due).

3/4/2009

CSE842, Spring 2009, MSU

2

Today's Topics

- Word Sense Disambiguation
- A quick introduction to Information Theory

3/4/2009

CSE842, Spring 2009, MSU

3

Word Sense Disambiguation

Words in Context

Sense	Examples (keyword in context)
1	... used to strain microscopic plant life from the ...
1	... too rapid growth of aquatic plant life in water ...
2	... automated manufacturing plant in Fremont ...
2	... discovered at a St. Louis plant manufacturing ...

Task: identify the correct sense based on the context

3/4/2009

CSE842, Spring 2009, MSU

4

Machine Learning Approaches

- Learn a classifier to assign one of possible word senses for each word
 - Acquire knowledge from labeled or unlabeled corpus
 - Human intervention only in labeling corpus and selecting set of features to use in training
- Input: feature vectors
 - Target
 - Context
- Output: learned models (e.g., classification rules) for unseen text

Input Features for WDS

- Word features
 - Word found in +/- k word window
 - Word immediately to the right (+1 W)
 - Word immediately to the left (-1 W)
 - Pair of words at offsets -2 and -1
 - Pair of words at offsets -1 and +1
 - Pair of words at offsets +1 and +2
- Part of speech and general classes (e.g., month)
 - Pair of tags at offsets -2 and -1
 - Tag at position -2, word at position -1
 - Etc.

Example

The ocean reflects the color of the sky, but even on cloudless days the color of the ocean is not a consistent blue. Phytoplankton, microscopic **plant** life that floats freely in the lighted surface waters, may alter the color of the water. When a great number of organisms are concentrated in an area, the plankton changes the color of the ocean surface. This is called a 'bloom.'

⇓

w_{-1} = Phytoplankton	t_{-1} = JJ
w_{+1} = life	t_{+1} = NN
w_{-2}, w_{-1} = (Phytoplankton, microscopic)	t_{-2}, t_{-1} = (NN, JJ)
w_{-1}, w_{+1} = (microscopic, life)	...
w_{+1}, w_{+2} = (life, that)	
word-within-k = ocean	
word-within-k = reflects	
word-within-k = color	
...	
word-within-k = bloom	

Supervised Learning

- Training and test sets with words labeled as to correct sense (*It was the biggest [fish: bass] I've seen.*)
 - Obtain feature vectors automatically (POS, co-occurrence information, etc.)
 - Run classifier on training data
 - Test on test data
 - Result: Classifier for use on unlabeled data

Naïve Bayesian Classifier

$$\hat{s} = \arg \max_{s \in S} P(s | FV) \quad \Rightarrow \quad \arg \max_{s \in S} \frac{p(FV|s)P(s)}{P(FV)}$$

↓

$$\arg \max_{s \in S} P(FV | s)P(s)$$

Independence Assumption $P(FV | s) \approx \prod_{j=1}^n P(f_j | s)$

$$\hat{s} = \arg \max_{s \in S} P(s) \prod_{j=1}^n P(f_j | s)$$

Decision List

Rule		Sense
<i>fish</i> within window	⇒	bass ¹
<i>striped bass</i>	⇒	bass ¹
<i>guitar</i> within window	⇒	bass ²
<i>bass player</i>	⇒	bass ²
<i>piano</i> within window	⇒	bass ²
<i>tenor</i> within window	⇒	bass ²
<i>sea bass</i>	⇒	bass ¹
<i>play/V bass</i>	⇒	bass ²
<i>river</i> within window	⇒	bass ¹
<i>violin</i> within window	⇒	bass ²
<i>salmon</i> within window	⇒	bass ¹
<i>on bass</i>	⇒	bass ²
<i>bass are</i>	⇒	bass ¹

Creating Decision List

A simple approach: Yarowsky '94's approach

- Every individual feature-value pair constitutes a test.
- These tests are then ordered according to their individual accuracy on entire training set based on log-likelihood ratio

$$\left| \text{Log} \left(\frac{P(\text{Sense}_1 | f_i)}{P(\text{Sense}_2 | f_i)} \right) \right|$$

Parameter Estimation

For each feature, we can get an estimate of conditional probability of sense 1 and sense2

Suppose the feature is: $w_{+1} = \textit{life}$

Count (sense 1 of plant, $w_{+1} = \textit{life}$) = 100

Count (sense 2 of plant, $w_{+1} = \textit{life}$) = 1

Maximum-likelihood estimate

$P(\text{sense 1 of plant} | w_{+1} = \textit{life}) = 100/101 = 0.99$

Any possible problem? (hw3)

A Partially Supervised Method

- A supervised method requires labeled data where data labeling can be very expensive
- Semi-supervised approach: use a small amount of labeled data and a large amount of unlabeled data

A Key Property: Redundancy

The ocean reflects the color of the sky, but even on cloudless days the color of the ocean is not a consistent blue. Phytoplankton, microscopic **plant** life that floats freely in the lighted surface waters, may alter the color of the water. When a great number of organisms are concentrated in an area, the plankton changes the color of the ocean surface. This is called a 'bloom.'

	⇓	
w_{-1} = Phytoplankton		word-within-k = ocean
w_{+1} = life		word-within-k = reflects
w_{-2}, w_{-1} = (Phytoplankton, microscopic)		word-within-k = bloom
w_{-1}, w_{+1} = (microscopic, life)		word-within-k = color
w_{+1}, w_{+2} = (life, that)		...

There are often many features which indicate the sense of the word

Heuristics

- One Sense per Collocation: certain words or phrases strongly associated with the target senses tend not to occur with the other sense.
- One Sense per Discourse: if the same word appears more than once in a document, then it is very likely to have the same sense every time.

Bootstrapping (Yarowsky'95)

- Start with a small set of seed examples: strongly associated with each sense (e.g. *sea* and *music* for **bass**), either intuitively or from corpus or from dictionary entries
- Gradually augment the seed examples with additional examples

Collecting Seed Examples

- Create a small subset of the labeled training data
 - Label a small number of examples by hand
 - Look through frequently occurring features, and label a few of them
 - Using words in dictionary definitions
 - Pick words in the two definitions for “plant”
- Partitioned the original collection into three sets:
 - 82 examples labeled with “life” sense
 - 106 examples labeled with “manufacturing” sense
 - 7350 unlabeled examples

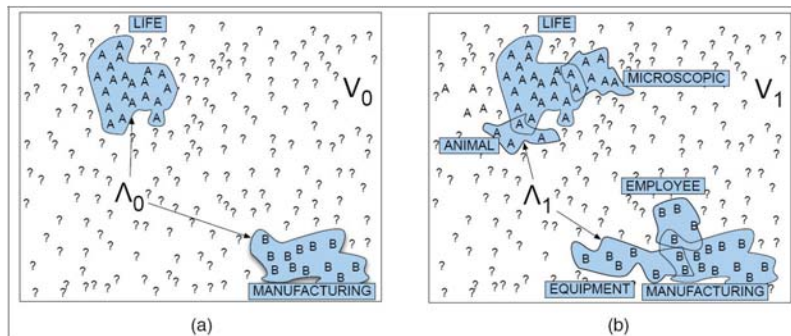
Bootstrapping

Train the supervised classification algorithm on seed examples.
 E.g., learn a decision list of all rules with weights above some threshold

LogL	Collocation	Sense
8.10	<i>plant life</i>	→ A
7.58	manufacturing plant	→ B
7.39	life (within ±2-10 words)	→ A
7.20	manufacturing (in ±2-10 words)	→ B
6.27	animal (within ±2-10 words)	→ A
4.70	equipment (within ±2-10 words)	→ B
4.39	employee (within ±2-10 words)	→ B
4.30	assembly <i>plant</i>	→ B
4.10	<i>plant closure</i>	→ B
3.52	<i>plant species</i>	→ A
3.48	automate (within ±2-10 words)	→ B
3.45	microscopic <i>plant</i>	→ A
	...	

Bootstrapping

Using the new rules to re-label the data (usually more data will be labeled this time)



Bootstrapping

- Use “one sense per discourse” constraint to filter and augment the tagged set.
- Repeat: induce a new set of rules with weight above the threshold from the labeled data

Stop when no new examples can be labeled

Labeling previously untagged contexts using the one-sense-per-discourse property

Change in tag	Disc. Num.	Training Examples (from same discourse)
A → A	724	... the existence of <i>plant</i> and animal life ...
A → A	724	... classified as either <i>plant</i> or animal ...
* → A	724	Although bacterial and <i>plant</i> cells are enclosed
A → A	348	... the life of the <i>plant</i> ; producing stem
A → A	348	... an aspect of <i>plant</i> life; for example
* → A	348	... tissues; because <i>plant</i> egg cells have
* → A	348	photosynthesis, and so <i>plant</i> growth is attuned

Error Correction using the one-sense-per-discourse property

Change in tag	Disc. Num.	Training Examples (from same discourse)
A → A	525	contains a varied <i>plant</i> and animal life
A → A	525	the most common <i>plant</i> life; the ...
A → A	525	slight within Arctic <i>plant</i> species ...
B → A	525	are protected by <i>plant</i> parts remaining from

Unsupervised Learning

- Cluster automatically derived feature vectors to ‘discover’ word senses using some similarity metric
 - Represent each cluster as average of feature vectors it contains
 - Label clusters by hand with known senses
 - Classify unseen instances by proximity to these known and labeled clusters
- Challenges:
 - What are the ‘right’ senses?
 - Cluster impurity
 - How do you know how many clusters to create?
 - Some clusters may not map to ‘known’ senses

A Quick Introduction to Information Theory

Definition of Information

- Information: *reduction in uncertainty*
- Let E be some event that occurs with probability P(E). If we are told that E has occurred, then we say we have received $I(E)=\log_2(1/P(E))$ bits of information
- Example:
 - Result of a fair coin flip ($\log_2 2=1$ bit)
 - Result of a fair die roll ($\log_2 6=2.585$ bits)
 - More information is provided by the outcome from a fair die roll than the outcome from a fair coin flip

Entropy

- Formula

$$H(p) = H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

- Notation

- **X**: A discrete set of symbols
- X : A random variable over **X**
- $p(x)$: The probability mass function of X

Explanation of Entropy

- The average uncertainty of a single random variable.
- The average amount of information in a random variable.
- The average number of bits required to transmit the outcome of the random variable.

More Explanation

- The average length of the message needed to transmit an outcome of that variable
- In general an optimal code sends a message of probability $p(i)$ in $\left\lceil \log \frac{1}{p(i)} \right\rceil$ bits
- Entropy: A weighted average of the code length:

$$H(X) = \sum_{x \in X} p(x) \log \frac{1}{p(x)} = E \left(\log \frac{1}{p(x)} \right)$$

Example

- Rolling an 8-sided die

$$H(X) = - \sum_{i=1}^8 p(i) \log p(i) = - \sum_{i=1}^8 \log \frac{1}{8} = \log 8 = 3 \text{ bits}$$

- Encoding the result as a 3 digit binary message:

1	2	3	4	5	6	7	8
001	010	011	100	101	110	111	000

Example: Simplified Polynesian

Small alphabet with the corresponding frequencies

p	t	k	a	i	u
1/8	1/4	1/8	1/4	1/8	1/8

Per-letter entropy

$$H(X) = - \sum_{i=\{p,t,k,a,i,u\}} p(i) \log p(i) = - \left[4 \times \frac{1}{8} \log \frac{1}{8} + 2 \times \frac{1}{4} \log \frac{1}{4} \right] = 2.5 \text{ bits}$$

p	t	k	a	i	u
100	00	101	01	110	111

Entropy

- Characteristics

- $H(X) \geq 0$
- $H(X) = 0$ only when the value of X is determinate
- Entropy increases with the message length (summable on independent random variables)

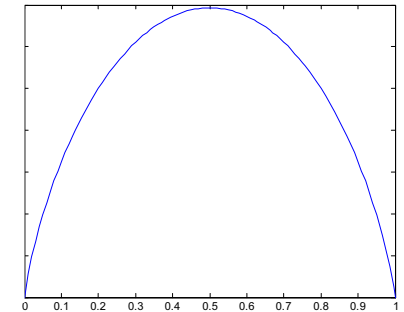
$$H(X_1 \cdots X_n) = H(X_1) + \cdots + H(X_n)$$

Entropy: two possible outcomes

$$H(P) = p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p}$$

Notice

- zero information at edges
- maximum information at 0.5 (1 bit)
- drop off more quickly close edges than in the middle



Joint Entropy

- Formula

$$H(X, Y) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(x, y)$$

- Interpretation

- The amount of information needed on average to specify values of both random variables

Joint Entropy

Two random variables: Temperature (T) and Humidity (M)

Joint Probability P(T, M)

	cold	mild	hot	
low	0.1	0.4	0.1	0.6
high	0.2	0.1	0.1	0.4
	0.3	0.5	0.2	1.0

T and M are not independent
 $P(T=t, M=m) \neq P(T=t)P(M=m)$

- $H(T) = 1.485, H(M) = 0.971$
- $H(T) + H(M) = 2.456$
- Joint Entropy
 - $H(T, M) = H(0.1, 0.4, 0.1, 0.2, 0.1, 0.1, 0.1) = 2.321$
 - $H(T, M) \leq H(T) + H(M)$

Conditional Entropy

- Formula

$$\begin{aligned}
 H(Y|X) &= \sum_{x \in X} p(x) H(Y|X=x) \\
 &= \sum_{x \in X} p(x) \left[- \sum_{y \in Y} p(y|x) \log p(y|x) \right] \\
 &= - \sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(y|x)
 \end{aligned}$$

- Interpretation

- The amount of extra information on average to communicate Y given known X

Conditional Entropy

- Conditional Entropy

- $H(T|M = \text{low}) = 1.252$
- $H(T|M = \text{high}) = 1.5$

Conditional Probability $P(T|M)$

	cold	mild	hot	
low	1/6	4/6	1/6	1.0
high	2/4	1/4	1/4	1.0

Average conditional entropy

$$\begin{aligned}
 P(T|M) &= \sum_m P(M=m) H(T|M=m) \\
 &= 0.6 \times 1.252 + 0.4 \times 1.5 = 1.351
 \end{aligned}$$

How much is M telling us on average about T ?

$$H(T) - H(T|M) = 1.485 - 1.351 = 0.134 \text{ bits}$$

Chain Rule for Entropy

$$\begin{aligned}
 H(X,Y) &= -E_{p(x,y)}(\log p(x,y)) \\
 &= -E_{p(x,y)}(\log p(x)p(y|x)) \\
 &= -E_{p(x,y)}(\log p(x) + \log p(y|x)) \\
 &= -E_{p(x)}(\log p(x)) - E_{p(x,y)}(\log p(y|x)) \\
 &= H(X) + H(Y|X) \\
 &= H(Y) + H(X|Y)
 \end{aligned}$$

Mutual Information

- Formula

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

- Interpretation

- The reduction in uncertainty of one random variable due to knowing about another
- The amount of information one random variable contains about another

- Characteristics

- Symmetric
- Non-negative
- Zero iff X, Y are independent

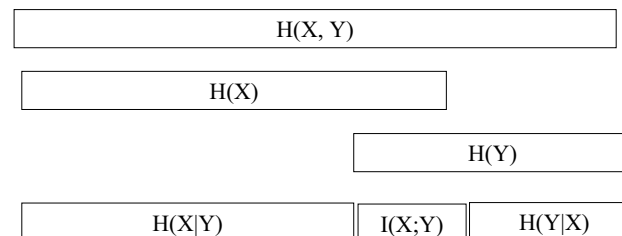
Mutual Information

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(X) + H(Y) - H(X,Y) \\ &= \sum_x p(x) \log \frac{1}{p(x)} + \sum_y p(y) \log \frac{1}{p(y)} + \sum_{x,y} p(x,y) \log p(x,y) \\ &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \end{aligned}$$

Pointwise mutual information (PMI)

$$I(x,y) = \log \frac{p(x,y)}{p(x)p(y)}$$

Relationship



K-L divergence

- Formular

$$D(p \parallel q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} = E_p \left(\log \frac{p(x)}{q(x)} \right)$$

- Interpretation

- A measure of how different two probability distributions are
- The average number of bits that are wasted by encoding events from a distribution p with a code based on a not-quite-right distribution q

- Characteristics

- Non-negative: $D(p \parallel q) = 0$ iff $p = q$
- Non-symmetric: $D(p \parallel q) \neq D(q \parallel p)$